FORM PTO-1390

U.S. DEPARTMENT OF COMMERCE PATENT AND TRADEMARK OFFICE
TRANSMITTAL LETTER TO THE UNITED STATES DESIGNATED/ELECTED OFFICE (DO/EO/US) CONCERNING A FILING UNDER 35 U.S.C. 371

ATTORNEY'S DOCKET NUMBER: 4001-1001

U.S. APPLN. NO. (If known, see 37 CFR 1.5): 10/031488

INTERNATIONAL APPLICATION NO.:	INTERNATIONAL FILING DATE:	PRIORITY DATE CLAIMED:
PCT/DE00/01791	31 MAY 2000 (31.05.00)	20 JULY 1999 (20.07.99)

TITLE OF INVENTION: METHOD FOR FORMING AND/OR UPDATING DICTIONARIES FOR THE AUTOMATIC READING OF ADDRESS

APPLICANT(S) FOR DO/EO/US: Thomas BAYER

Applicant herewith submits to the United States Designated/Elected Office (DO/EO/US) the following items and other information:

1. [X] This is a **FIRST** submission of items concerning a filing under 35 U.S.C. 371.
2. [ ] This is a **SECOND** or **SUBSEQUENT** submission of items concerning a filing under 35 U.S.C. 371.
3. [X] This express request to begin national examination procedures (35 U.S.C. 371(f)) at any time rather than delay examination until the expiration of the applicable time limit set in 35 U.S.C. 371(b) and PCT Articles 22 and 39(1).
4. [X] A proper Demand for International Preliminary Examination was made by the 19th month from the earliest claimed priority date.
5. [X] A copy of the International Application as filed (35 U.S.C. 371(c)(2))
   a. [X] is transmitted herewith (required only if not transmitted by the International Bureau).
   b. [ ] has been transmitted by the International Bureau. (see attached copy of PCT/IB/308)
   c. [ ] is not required, as the application was filed in the United States Receiving Office (RO/US).
6. [X] A translation of the International Application into English (35 U.S.C. 371(c)(2)).
7. [ ] Amendments to the claims of the International Application under PCT Article 19 (35 U.S.C. 371(c)(3)).
   a. [ ] are transmitted herewith (required only if not transmitted by the International Bureau).
   b. [ ] have been transmitted by the International Bureau.
   c. [ ] have not been made; however, the time limit for making such amendments has NOT expired.
   d. [ ] have not been made and will not be made.
8. [ ] A translation of the amendments to the claims under PCT Article 19 (35 U.S.C. 371(c)(3)).
9. [X] An oath or declaration of the inventor(s) (35 U.S.C. 371(c)(4)).
10. [ ] A translation of the annexes of the International Preliminary Examination Report under PCT Article 36 (35 U.S.C. 371(c)(5)).

Item 11. to 16. below concern document(s) or information included:

11. [X] An Information Disclosure Statement under 37 CFR 1.97 and 1.98.
12. [X] An assignment document for recording. A separate cover sheet in compliance with 37 CFR 3.28 and 3.31 is included.
13. [X] A **FIRST** preliminary amendment.
    [ ] A **SECOND** or **SUBSEQUENT** preliminary amendment.
14. [ ] A substitute specification.
15. [ ] A change of power of attorney and/or address letter.
16. [X] Other items or information: INTERNATIONAL PRELIMINARY EXAMINATION REPORT (PCT/IPEA/409), INTERNATIONAL SEARCH REPORT (PCT/ISA/210), APPLICATION DATA SHEET, ABSTRACT

U.S. APPLICATION NO. (if known, see 37 CFR 1.5)	INTERNATIONAL APPLICATION NO.	ATTORNEY'S DOCKET NO.
10/031488	PCT/DE00/01791	4001-1001

17. [X] The following fees are submitted:

	CALCULATIONS PTO USE ONLY
BASIC NATIONAL FEE (37 CFR 1.492(a)(1)-(5)): Neither international preliminary examination fee (37 CFR1.482) nor international search fee (37 CFR1.445(a)(2)) paid to USPTO and International Search Report not prepared by the EPO or JPO ..... $ 1,040.00	
International preliminary examination fee (37 CFR 1.482) not paid to USPTO but International Search Report prepared by the EPO or JPO ..... $ 890.00	
International preliminary examination fee (37 CFR 1.482) not paid to USPTO but international search fee (37 CFR 1.445(a)(2)) paid to USPTO ..... $ 740.00	
International preliminary examination fee (37 CFR 1.482) paid to USPTO but all claims did not satisfy provisions of PCT Article 33(1)-(4) ..... $ 710.00	
International preliminary examination fee (37 CFR 1.482) paid to USPTO and all claims satisfied provisions of PCT Article 33(1)-(4) ..... $ 100.00	
ENTER APPROPRIATE BASIC FEE AMOUNT =	$ 890.00
Surcharge of $130.00 for furnishing the oath or declaration later than months from the earliest claimed priority date (37 CFR 1.492(e)).	$

CLAIMS	NUMBER FILED	NUMBER EXTRA	RATE	$
Total claims	11 - 20 =	0	X $18.00	$
Independent claims	1 - 3 =	0	X $84.00	$
MULTIPLE DEPENDENT CLAIMS(S) (if applicable)			+ $280.00	$

TOTAL OF ABOVE CALCULATIONS =	$ 890.00
Reduction of ½ for filing by small entity, if applicable. Applicant claims Small Entity Status under 37 CFR 1.27. +	$
SUBTOTAL =	$ 890.00
Processing fee of $130 for furnishing the English translation later than months from the earliest claimed priority date (37 CFR1.492(f)).	$
TOTAL NATIONAL FEE =	$ 890.00
Fee for recording the enclosed assignment (37 CFR1.21(h)). The assignment must be accompanied by an appropriate cover sheet (37 CFR 3.28, 3.31). $40.00 per property +	$ 40.00
TOTAL FEES ENCLOSED =	$ 930.00
Amount to be refunded:	
charged:	

a. [X] A check in the amount of $ **930.00** to cover the above fees is enclosed.

b. [ ] Please charge my Deposit Account No. **25-0120** in the amount of $ to cover the above fees. A duplicate copy of this sheet is enclosed.

c. [X] The Commissioner is hereby authorized to charge any additional fees which may be required by 37 CFR 1.16 and 1.17, or credit any overpayment to Deposit Account No. **25-0120**. A duplicate copy of this sheet is enclosed.

SEND ALL CORRESPONDENCE TO:

YOUNG & THOMPSON
745 South 23rd Street
2nd Floor
Arlington, VA 22202
(703) 521-2297
facsimile (703) 685-0573
Customer Number: 000466

January 22, 2002

By Benoît Castel
Benoît Castel
Attorney for Applicant
Registration No. 35,041

PATENTS

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of

Thomas BAYER

Serial No. (unknown)

Filed herewith

METHOD FOR FORMING AND/OR UPDATING DICTIONARIES FOR THE AUTOMATIC READING OF ADDRESSES

PRELIMINARY AMENDMENT

Commissioner for Patents

Washington, D.C. 20231

Sir:

Prior to the first Official Action and calculation of the filing fee, please substitute Claim 1 as originally filed, which appears on page 16, with new Claim 1 as filed in the Article 34 amendment of July 5, 2001. The page containing Claim 1 is marked "AMENDED SHEET" and is attached hereto. Following the insertion of Claim 1, please amend these claims as follows:

IN THE CLAIMS:

Please amend the following claims:

--8.(Amended) The method as claimed in claim 1, characterized in that for word groups having n words, n > 1, the words having a distance from one another of m words, m > = 0, the addresses are searched with windows having a width of n + m words starting with the respective single word determined for the dictionary and when further n-1 single words determined for the dictionary

have been found in the predetermined gaps m between one another, these word groups found are included with their frequencies in the corresponding dictionary.

9.(Amended) The method as claimed in claim 1, characterized in that the factor of similarity between the words is determined by means of the Levenshtein method.

10.(Amended) The method as claimed in claim 1, characterized in that the dictionary entries to be removed and the new entries in the dictionary are displayed, categorized and confirmed at a video coding station.

11.(Amended) The method as claimed in claim 1, characterized in that the words and/or word groups to be entered into the dictionary, before they are entered, are compared with the contents of a file in which generally valid names characteristic of the respective dictionary category, or at least character strings, are stored and are transferred into the corresponding dictionary if they correspond.--

IN THE ABSTRACT:

Please delete the abstract as originally filed which appears on page 19. Add new abstract as enclosed herewith on a separate sheet.

R E M A R K S

The above changes in the abstract and claims merely place this national phase application in the same condition as it was during Chapter II of the international phase, with the multiple dependencies being removed. Following entry of this amendment by substitution of the pages, only claims 1-11 remain pending in this application. Attached hereto is a marked-up version of the changes made to the abstract and claims by the current amendment. The attached page is captioned **"VERSION WITH MARKINGS TO SHOW CHANGES MADE"**.

Respectfully submitted,

YOUNG & THOMPSON

By Benoît Castel

Benoît Castel
Attorney for Applicant
Customer No. 000466
Registration No. 35,041
745 South 23rd Street
Arlington, VA 22202
703/521-2297

January 22, 2002

"VERSION WITH MARKINGS TO SHOW CHANGES MADE"

Claims 8-11 have been amended as follows:

8. (Amended) The method as claimed in ~~one of claims 1 and 2,~~ claim 1, characterized in that for word groups having n words, n > 1, the words having a distance from one another of m words, m > = 0, the addresses are searched with windows having a width of n + m words starting with the respective single word determined for the dictionary and when further n-1 single words determined for the dictionary have been found in the predetermined gaps m between one another, these word groups found are included with their frequencies in the corresponding dictionary.

9. (Amended) The method as claimed in ~~one of claims 1, 2, 7, 8,~~ claim 1, characterized in that the factor of similarity between the words is determined by means of the Levenshtein method.

10. (Amended) The method as claimed in ~~one of claims 1 to 9,~~ claim 1, characterized in that the dictionary entries to be removed and the new entries in the dictionary are displayed, categorized and confirmed at a video coding station.

11. (Amended) The method as claimed in ~~one of claims 1 to 9,~~ claim 1, characterized in that the words and/or word groups to be entered into the dictionary, before they are entered, are compared with the contents of a file in

which generally valid names characteristic of the respective dictionary category, or at least character strings, are stored and are transferred into the corresponding dictionary if they correspond.

The abstract has been amended as follows:

Abstract

Method for forming and/or updating dictionaries for the automatic reading of addresses

The reading results of an agreed number of images of items, achieved by the OCR reader, are temporarily stored subdivided into reading results which are read unambiguously and reading results which are~~rejected.~~

~~rejected.~~ Then classes of words or word groups belonging together of the reading results temporarily stored and rejected, consisting in each case of n address words, n = 1,2,...,a, with interword gaps m, m = 0,1,...b are formed which do not drop below a particular similarity factor referred to in each case a particular n and m value between them. In the dictionary or dictionaries of the associated address areas, representatives, at least, of the classes whose frequency exceeds a predetermined value are included.

~~Figure 1~~

1. Method for forming and/or updating dictionaries for automatic reading of addresses,
characterized by the following steps:

- buffering of the reading results achieved by the OCR reader, i.e. the results of the addresses of an established number of transmission images or transmission images read within an established time interval, divided into unambiguously read results with agreement with the dictionary entry and into rejected reading results without agreement with the dictionary entry,

- formation of classes of words or associated word groups with the pertinent representatives of the buffered and rejected reading results, the word groups consisting of n address words n = 1, 2,...a, between which m, m = 0, 1, ...b, additional words at a time are located, and the words of the classes of words or the words of the classes of word groups, relative to a certain n-value and m-value at the time, among one another do not fall below a certain similarity quantity,

- acceptance of at least one representative of those classes with a frequency which exceeds a fixed value into the dictionary or dictionaries of the assigned address areas.

Abstract

Method for forming and/or updating dictionaries for the automatic reading of addresses

The reading results of an agreed number of images of items, achieved by the OCR reader, are temporarily stored subdivided into reading results which are read unambiguously and reading results which are rejected. Then classes of words or word groups belonging together of the reading results temporarily stored and rejected, consisting in each case of n address words, n = 1,2,...,a, with interword gaps m, m = 0,1,...b are formed which do not drop below a particular similarity factor referred to in each case a particular n and m value between them. In the dictionary or dictionaries of the associated address areas, representatives, at least, of the classes whose frequency exceeds a predetermined value are included.

Description

Method for forming and/or updating dictionaries for the automatic reading of addresses

The invention relates to a method for forming and/or updating dictionaries for reading addresses.

Address reading systems need information on the content and syntax of addresses in order to be able to extract the required information such as town, zip code, first and last name, etc. The permissible content of individual address elements is described by means of a dictionary (list of permissible strings) which, according to the prior art, is built up from present information sources such as, e.g. from a postal dictionary or from a list of employees of a company. However, the application domain changes with time so that the dictionary created at the beginning no longer completely includes all existing contents. It is especially when a reading system is used for mail distribution within a company, that the change in the set of words is considerable: employees leave the company, new employees are added, employees change their department or last names due to marriage, etc. Thus, entries are missing in the dictionary and there are entries which are no longer valid. The more the set of words currently used deviates from the lexicon, the more the recognition performance of the reading system drops.

Previously, these changes had to be manually transferred into the dictionaries at certain time intervals so that the disadvantages described occurred.

It is the object of the invention to automatically form and/or automatically update a dictionary for reading addresses.

According to the invention, the object is achieved by the features of claim 1. This is based on the concept of temporarily storing the results of the current reading processes, to evaluate them and to use them for automatically building up or updating a dictionary. During the temporary storage, the respective address is marked to indicate whether it has been read successfully or whether it has been rejected. If a dictionary is to be newly created or if new addressees are to be entered in the existing dictionary, the rejected reading results are utilized.

The dictionaries can contain individual words, e.g. last names and/or coherent word groups with words, etc. first and last name or first and last name and street names, where the words are located both directly next to one another (gap m = 0) and can also be spaced apart by m words.

Automatic building up of a dictionary or, respectively, automatic updating of the dictionary due to new addressees or changes in the addressees is possible by forming classes of words or word groups which have a fixed minimum measure of similarity with respect to one another, and including at least the representative in the dictionary or dictionaries of the associated address areas.

Advantageous embodiments of the invention are described in the subclaims.

To form classes, it is advantageous to create a list of all words/word groups of the rejected reading results which are sorted in accordance with the frequency of the words/word groups. Beginning with the most frequent

word/word group, the factor of similarity with all remaining words/word groups is determined and entered in a similarity list. All words/word groups in the similarity list having a similarity factor above a fixed threshold are then allocated as class to the current word/word group. After that, the words/word groups of the class formed are removed from the frequency list. The representatives of the respective class of words or word groups of the reading results temporarily stored and rejected can be formed by the shortest or most frequent word or word groups.

To recognize addresses in the dictionary which must be changed or removed, it is advantageous to statistically analyze the addresses read unambiguously. If there is an abrupt change in the frequency of words and/or word groups beyond a particular threshold and if it persists for a predetermined time, these words/word groups are removed from the dictionary.

To avoid irrelevant words of the reading results from being included in the dictionary, they can be determined by comparison with words stored in a special file for irrelevant words.

It is also of advantage in this connection not to include short words of less than p letters and without fullstop as irrelevant in the dictionary. To perform the address interpretation in as detailed as manner as possible with the aid of the dictionaries, it is advantageous to include, in addition to the representatives, also the words and/or word groups of the associated classes with the similarity factors and frequencies.

In a further advantageous embodiment, word groups belonging together and having n words which are mutually spaced apart by m words can be determined in that the addresses are searched with windows having a

width of n + m words starting with the respective individual word determined for the dictionary. Once the further n - 1 individual words with the gaps of m words between them have been determined, this word group and its frequencies are included in the corresponding dictionary.

It is also advantageous to determine the similarity factor by means of the Levenshtein method (see "A Method for the Correction of Garbled Words, based on the Levenshtein Metric", K. Okuda, E. Tanaka, T. Kasai, IEEE Transactions on Computers, Vol. c-25, No. 2, February 1976).

It can also be advantageous to categorize, and to have confirmed, the dictionary updatings found at a video coding station or to compare the new entries into the dictionary additionally, before they are taken into the corresponding category, with the contents of a file in which characteristic generally applicable names or at least strings related to the respective category (first name, last name, department) are stored.

In the text which follows, the invention will be explained in greater detail in an exemplary embodiment and referring to the drawing. The aim is to determine previously unknown last names (n = 1) or pairs of unknown first and last names (n = 2) or last and/or first and last names and department names of employees of a company and/or corresponding, no longer valid names or name combinations, and to perform dictionary changes.

figure 1 shows a flow structure of a monitor process for monitoring and controlling the updating of the dictionary

figure 2 shows a flow structure for determining and marking a relevant words

figure 3 shows a flow structure for determining previously unknown single words (n = 1) (last names)

figure 4 shows a flow structure for determining previously unknown word groups starting with the single words

figure 5 shows a flow structure for updating the dictionaries, taking into consideration the word categories.

The word proposals are automatically generated from the recognition results calculated for each pattern of an item by the reading system in daily operation. The recognition results for each pattern of an item comprise different geometric objects (layout objects) such as text blocks, lines, words and characters and their relations to one another, that is to say which lines belong to which text block, which words are located in which lines etc. For each individual character pattern, the reading system generates a list of possible character meanings. In addition, the reading system calculates for each layout object its position in the pattern of an item and its geometric dimensions.

To update or even learn dictionary entries, the set of items processed is separated into two subsets, into the set of items read automatically (but not necessarily correctly) by the reading system and the set of rejected items. The set of items read automatically is used for determining dictionary entries which are no longer valid; from the set of rejected items, new dictionary entries are derived.

The exemplary system consists of five modules: a monitor process, processing of the recognition results (preprocessing), two dictionary generation methods and a proposal administrator.

The monitoring process according to figure 1 monitors and controls the dictionary training. The recognition results 21 for each pattern of an item, together with an identification for "read successfully" or "rejected", are transferred from the reader to the monitor. Additional information on the type of item (letter, large letter, in-house mail form) and other features relating to the individual objects of the recognition results such as ROI (Region of Interest), line and word hypotheses, disassembly alternatives and character recognition results can also be transferred. These recognition results are stored in a buffer 22 in the monitor until a sufficiently large amount of data has accumulated (e.g. after 20 000 items or after one week of operation).

In the simplest case, only the first alternative of the character recognition results together with the best segmenting path is stored in a buffer. For example, the content could look as follows:

<Recognition results>	<Identification>
:...	
1017921 PMD 55	recognized
MR. ALFRED C SCHMIDI	
EXCCU1LVE DIRCC1OR, 0PCRA1IONS	
DCVC1OPMENT	
MyComp, INC	
1 MyStreet	
MyCity, 12345	
PO11Y O/BRIEN	rejected,
	not in the dictionary
MANAGER, COMMUNITY AFFAIRS	
MyComp INC	
1 MyStreet	
MyCity, 12345	

PO1LY OBRIEN MANAGER, COMMUNITY AFFAIRS MyComp, INC 1 MyStreet MyCity, 12345	rejected not in the dictionary
MS ME1INDA DUCKSWORTH MyComp, INC MAI1 CODE 63-33 1 MyStreet MyCity, 12345	recognized
*********AURO**MIXED AADC 460 MIK0 SCHWARTZ O AND T 26-00 1 MyStreet MyCity, 12345	Rejected, not in the dictionary

...

If sufficient results are available, the rejected recognition results are transferred to a processing unit 30 and forwarded to the two subprocesses for dictionary training for single words 50 and word groups 60. In the case of a successful automatic recognition, the results are transferred to a statistics module 40. When all items have been processed, the word and word group lists 41 of the statistics module and of the dictionary training processes 51, 61 are collected and presented to an operator for confirmation by means of a suitable graphical user interface.

In the processing unit 30, irrelevant words in the rejected recognition results are identified which are

not taken into consideration in the subsequent text analysis (compare figure 2). These words are marked as not relevant but are not deleted since the word neighborhood is of importance for the subsequent building up of the dictionary.

In the method step marking irrelevant words 31, short words are marked from the set of word hypotheses, for example those words which are less than 4 letters long and, at the same time, do not have a fullstop, and those, less than 50% of whose characters are alphanumeric. Furthermore, those words are marked which are contained in a special file 32 which contains frequent but irrelevant words for this application. In the application of in-house mail distribution, for example, this special lexicon can contain the company name, city name, street name, post box designation etc. The results of the processing are written back into a buffer 33.

After the preprocessing, the results look as follows:

<title MR> <first-name ALFRED> <last-name SCHMID>
<role EXECUTIVE DIRECTOR OPERATIONS>

PO11Y O/BRIEN
MANAGER, COMMUNITY AFFAIRS
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

PO1LY OBRIEN
MANAGER, COMMUNITY AFFAIRS
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

<title MS> <first-name MELINDA> <last-name DUCKSWORTH>

<non-alpha ********AURO**MIXED> AADC <short 460>
MIKO SCHWARTZ
<short O> <short AND> <short T> 26-00
<irrelevant MyComp, INC>
<irrelevant 1 MyStreet>
<irrelevant MyCity> <irrelevant 12345>

...

According to figure 3, from the processed rejected recognition results, a frequency list FL 53 of all words occurring there is created in first step 52, sorted in accordance with descending frequency and stored in a buffer. For the above example, the frequency list FL 53 could look as follows:

...	
AFFAIRS	37
MANAGER	37
COMMUNITY	37
OBRIEN	20
O/BRIEN	17
SCHWARTZ	15
MIKO	12
POLLY	10
PO11Y	8
PAULA	8
PO1LY	5
MIKO	3
...	

From this list, a dictionary W1 of relevant words 51 is built up step by step. For each word in the frequency list FL 53, the distance d to all words in this frequency list is determined. One method for measuring the distance between two strings is the Levenshtein method which calculates the minimum distance between

two strings referred to 3 cost categories, at the cost of replacing one character, an insertion and a deletion operation. In addition to the string, other features of the recognition result, for example the character alternatives, the segmentation alternatives, etc., can be used for calculating d.

The first word in the frequency list FL 53 (the currently most frequent one) is included in the dictionary W1 51 and deleted 54 from the frequency list FL 53. All words from the frequency list FL 53 having a distance of less than a predetermined threshold th_d are allocated 55, 56 to the current word in the dictionary W1 51 with their frequency. At the same time, these words are deleted in the frequency list FL 53. The iteration stops when the frequency list FL 53 is empty. This forms word classes which do not exceed a distance d between each other or, respectively, do not drop below a corresponding similarity factor.

When all words have been processed, the dictionary W1 51 consists of a set of word classes. The shortest word of a word class is called the representative of the group. Each word class contains words which are similar to each other, with the associated frequencies and distances from the class representative. The representatives of word classes in the dictionary W1 51, and thus also the word classes, are sorted 57 in accordance with descending frequency. The frequency of a word class is composed of the frequency of the representative and the frequencies of the elements of the word class. Word classes with a frequency which drops below a particular threshold are deleted from the dictionary W1 51. In consequence, the following dictionary W1 51 is formed from the above list:

==========

<Word class>	<Frequency>	<Distance>
...		
AFFAIRS	37	
MANAGER	37	
COMMUNITY	37	
OBRIEN	37	
O/BRIEN	17	(d = 1)
POLLY	23	
PO11Y	8	(d = 2)
PO11Y	5	(d = 1)
SCHWARTZ	15	
MIK0	15	
MIKO	3	(d = 1)
PAULA	8	
...		

The formation of representatives can be supported with further knowledge depending on the application. Thus, a word can be mapped either onto a number or onto an alpha sequence by using OCR replacement tables which define interchangeable pairs of characters such as 1 - L, 0 - 0, 2 - Z, 6 - G etc. If, in addition, alternative sets for word classes to be learnt are known, for example nicknames for first names such as Paula-Polly, Thomas-Tom, etc., this replacement can also be performed. Both steps can be applied to the dictionary W1 51 which leads to a further blending of word classes.

Finally, all words occurring in the dictionary W1 51 are marked in the recognition results and supplemented by their representative. In the text which follows these words will be called W1 words.

At the top of the dictionary W1 51, the most frequent, previously unknown word forms are located and the word classes contain spelling variants thereof. Thus, in the application of in-house mail distribution, previously unknown first and second names and parts of

departmental designations will be in the dictionary W1 51. In addition, their word classes contain spelling variants or variants which have arisen due to the characteristics of the reading system.

Starting with the representatives of the word classes in the dictionary W1 51 which are marked as such in the recognition results, word groups of length 2 to n are determined in the next step according to figure 4 in that the neighborhoods of W1 words of the recognition results 62 are examined. For each W1 word, the right-hand neighborhood is searched in a window of width $k <= n$ to see whether it contains further W1 words. n-1 initially empty dictionaries are set up in a buffer and filled step by step. An n-tuple is then included in a word group buffer 53 when n W1 words have been found and there are fewer than m further non-W1 words between these n. As in the case of the dictionary W1 51, the frequency of occurrence of the individual word groups of length n is stored here, too.

The choice of the values of m and n depends on the actual application. For values of $n > 4$, no further significant frequent entries can be expected in the application of reading addresses. $m = 0$ means that all n W1 words follow one another directly. In the case of pairs of first and last names, however, in particular, a second name can occasionally interrupt the direct succession, just as segmentation errors of the automatic reader can generate supposed word hypotheses and thus prevent a direct succession. In consequence, $m = 1$ and $n = 3$ are suitable values for the application described. In this step, in consequence, $n - 1$ dictionaries Wn 61 containing frequent word sequences with the frequencies for pairs, triplets etc. up to n-tuple are generated from the word group buffer. In each dictionary Wn 61, the frequencies of the n-tuples are included with the frequencies of the W1 words of the n-tuples to calculate a dimension. Each dictionary Wn 61

is sorted in accordance with descending dimensions so that the most significant word groups are again at the beginning of each dictionary Wn 54.

For the above example, the dictionary W2 looks as follows:

W2

COMMUNITY AFFAIRS	37
MANAGER COMMUNITY	37
POLLY OBRIEN	23
MIK0 SCHWARTZ	15
PAUL OBRIEN	8

The dictionary W3 has 3 entries provided that the name POLLY OBRIEN always occurs in combination with the designation MANAGER COMMUNITY AFFAIRS and that a line break is allowed in an n-tuple.:

W3

MANAGER COMMUNITY AFFAIRS	37
POLLY OBRIEN MANAGER	23
OBRIEN MANAGER COMMUNITY	23

As described, the word proposals of the dictionaries Wn 61 (W2, W3, etc.) are now presented to an operator for validation according to figure 5. Knowledge about the word units 72 to be learnt makes it possible at this point to categorize 71 entries in the dictionaries W1, W2, ... Wn 51, 61 semantically. Thus, in this application, entries can be allocated to the semantic class <Name> by looking at generally applicable lists of first names. This similarly applies to the semantic class <Department> which can be derived from keywords such as Department. Naturally, this process can also

be carried out automatically without an operator by comparison with the entries of these lists.

For items successfully distributed, the address elements required for this have been found and are identified as such in the recognition results. If, for example, last name and first name have been successfully read in the application of the in-house mail distribution, these results are registered in statistics; in particular, the frequency of the extracted words, pairs, generally of in-tuples over defined time intervals td, e.g. for a week, are stored and it is possible to take into consideration the type of item. As a result, a distribution of the address elements to be extracted for a sequence of time intervals is obtained:

```
==========================================================
Time 1

MELINDA DUCKSWORTH       123
ALFRED SCHMID                      67
...

Time 2
MELINDA DUCKSWORTH         1
ALFRED SCHMID                      85
...

Time 3
MELINDA DUCKSWORTH         2
ALFRED SCHMID                      72
...
==========================================================
```

From the distribution thus found, it is possible to derive whether dictionary entries are to be deleted: the entries are inserted into a list for removal from the dictionary if their frequency abruptly decreases

from td_i to td_{i+1} and stays at this level in successive time intervals td_{i+k} (e.g. k = 4). Thus, the person MELINDA DUCKSWORTH in the above example is deleted from the dictionary. This sequence can also be additionally conducted via a confirmation process.

Patent claims

1. A method for forming and/or updating dictionaries for the automatic reading of addresses, characterized by the following steps:
- temporary storage of the reading results, achieved by the OCR reader, of the addresses of an agreed number of images of items or of images of items read within an agreed period of time, subdivided into unambiguously read results with correspondence with a dictionary entry and into rejected reading results without correspondence with a dictionary entry,
- formation of classes of words with associated representatives or word groups, belonging together, of the temporarily stored and rejected reading results, consisting in each case of n address words, n = 1, 2, ...a, with interword gaps m, n = 0,1,...b which do not drop below a particular similarity factor referred to a particular n value and m value in each case,
- inclusion of at least the representatives of the classes whose frequency exceeds a predetermined value, into the dictionary or dictionaries of the associated address areas.

2. The method as claimed in claim 1, characterized in that
- for the purpose of forming classes, a frequency list of all words or word groups of the rejected reading results occurring is created, sorted in accordance with their frequency,
- for each word or each word group, beginning with the most frequent word or the most frequent word group, the factor of similarity with all remaining words or word groups is determined and entered in a similarity list,

- all words or word groups in the similarity list with a similarity factor above a predetermined threshold are allocated as a class to the current word or the current word group,
- subsequently the words or word groups of the class formed in each case are removed from the frequency list.

3. The method as claimed in claim 1, characterized in that the representative of the respective class of words or word groups of the reading results temporarily stored and rejected is formed by the shortest or most frequent word or word group.

4. The method as claimed in claim 1, characterized in that the temporal frequency of the words or word groups of the addresses read unambiguously is statistically analyzed with the aim of removing the respective entered words or word groups from the dictionary in the case of their abrupt reduction, lasting over a predetermined period of time, over a predetermined threshold.

5. The method as claimed in claim 1, characterized in that the irrelevant words of the reading results are determined by comparison with words stored in a special file and are not included in the dictionary.

6. The method as claimed in claim 1, characterized in that short words having fewer than p letters and without fullstop are not included in the dictionary.

7. The method as claimed in claim 1, characterized in that in addition to the representatives, the words and/or word groups of the associated classes with

the similarity factors and frequencies are entered in the dictionary.

8. The method as claimed in one of claims 1 and 2, characterized in that for word groups having n words, n > 1, the words having a distance from one another of m words, m > = 0, the addresses are searched with windows having a width of n + m words starting with the respective single word determined for the dictionary and when further n-1 single words determined for the dictionary have been found in the predetermined gaps m between one another, these word groups found are included with their frequencies in the corresponding dictionary.

9. The method as claimed in one of claims 1, 2, 7, 8, characterized in that the factor of similarity between the words is determined by means of the Levenshtein method.

10. The method as claimed in one of claims 1 to 9, characterized in that the dictionary entries to be removed and the new entries in the dictionary are displayed, categorized and confirmed at a video coding station.

11. The method as claimed in one of claims 1 to 9, characterized in that the words and/or word groups to be entered into the dictionary, before they are entered, are compared with the contents of a file in which generally valid names characteristic of the respective dictionary category, or at least character strings, are stored and are transferred into the corresponding dictionary if they correspond.

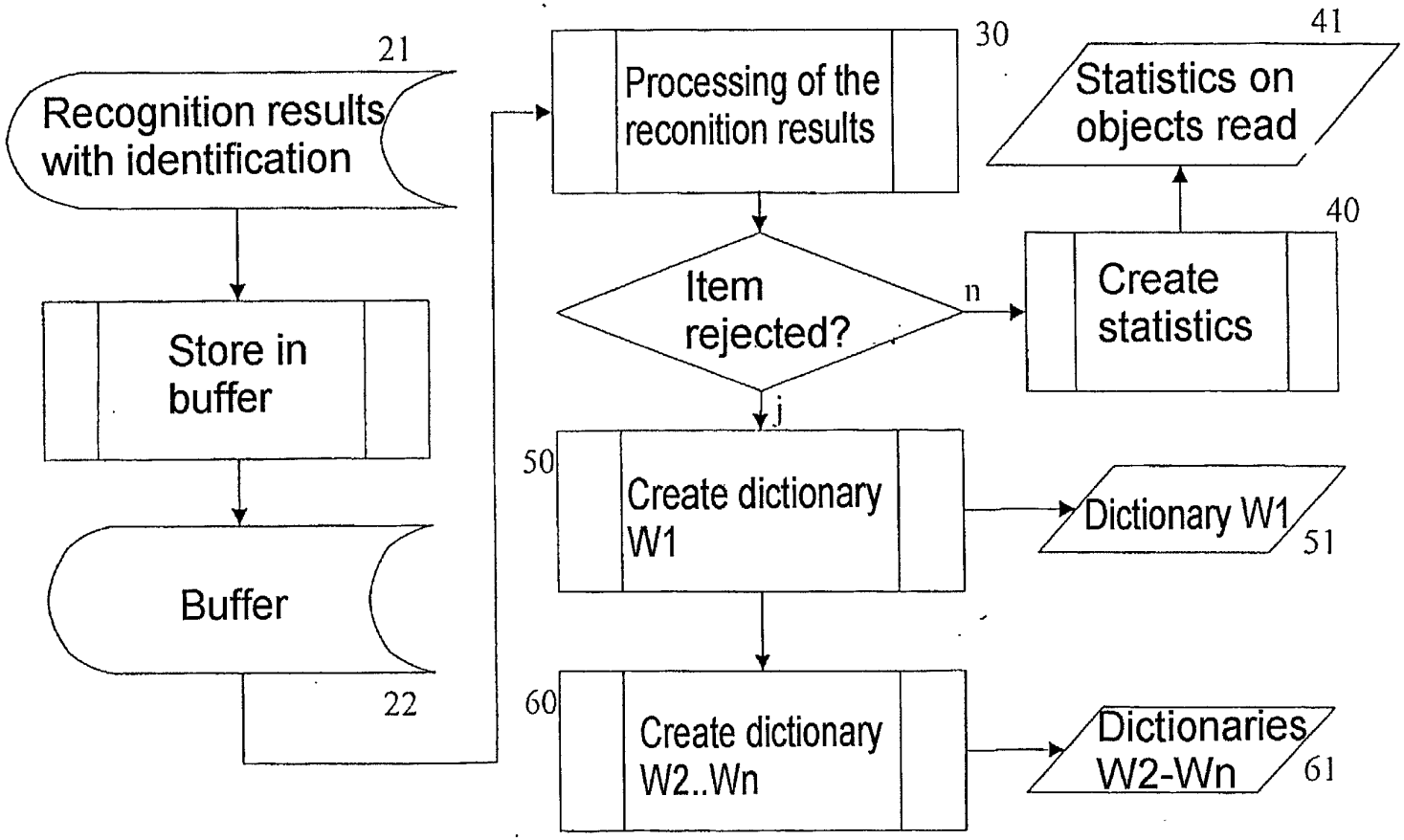

21
Recognition results with identification
Store in buffer
Buffer
22
30
Processing of the reconition results
Item rejected?
n
j
40
Create statistics
41
Statistics on objects read
50
Create dictionary W1
Dictionary W1
51
60
Create dictionary W2..Wn
Dictionaries W2-Wn
61

FIG 1

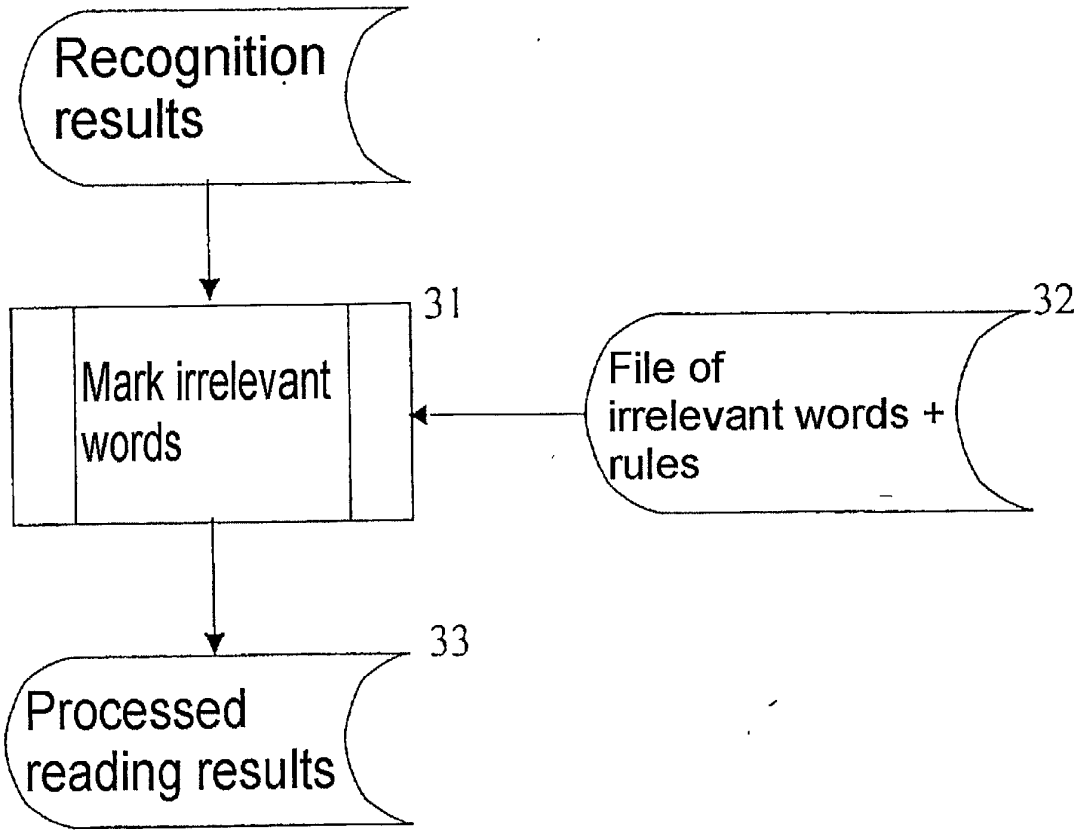
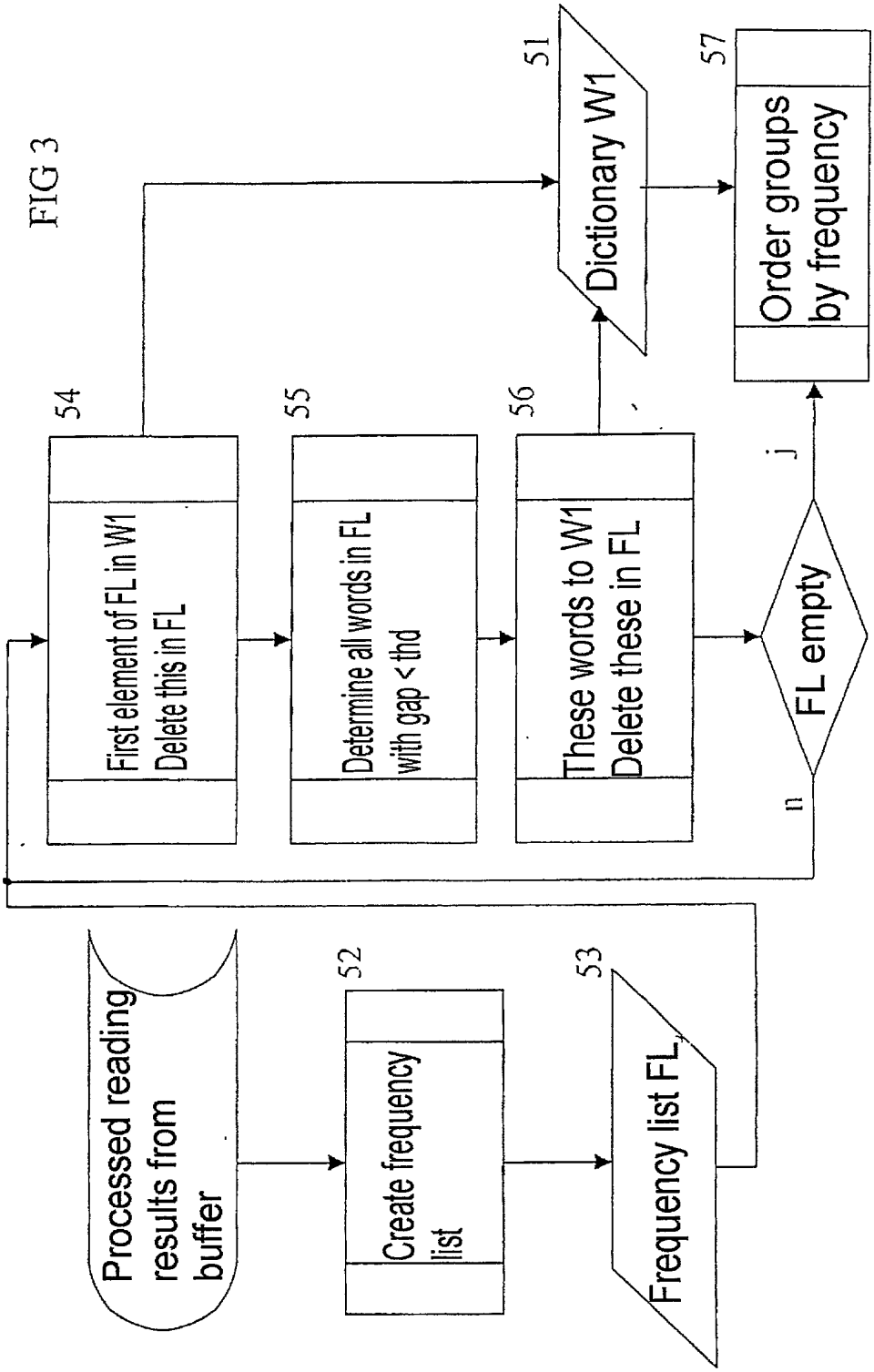


FIG 2

FIG 3



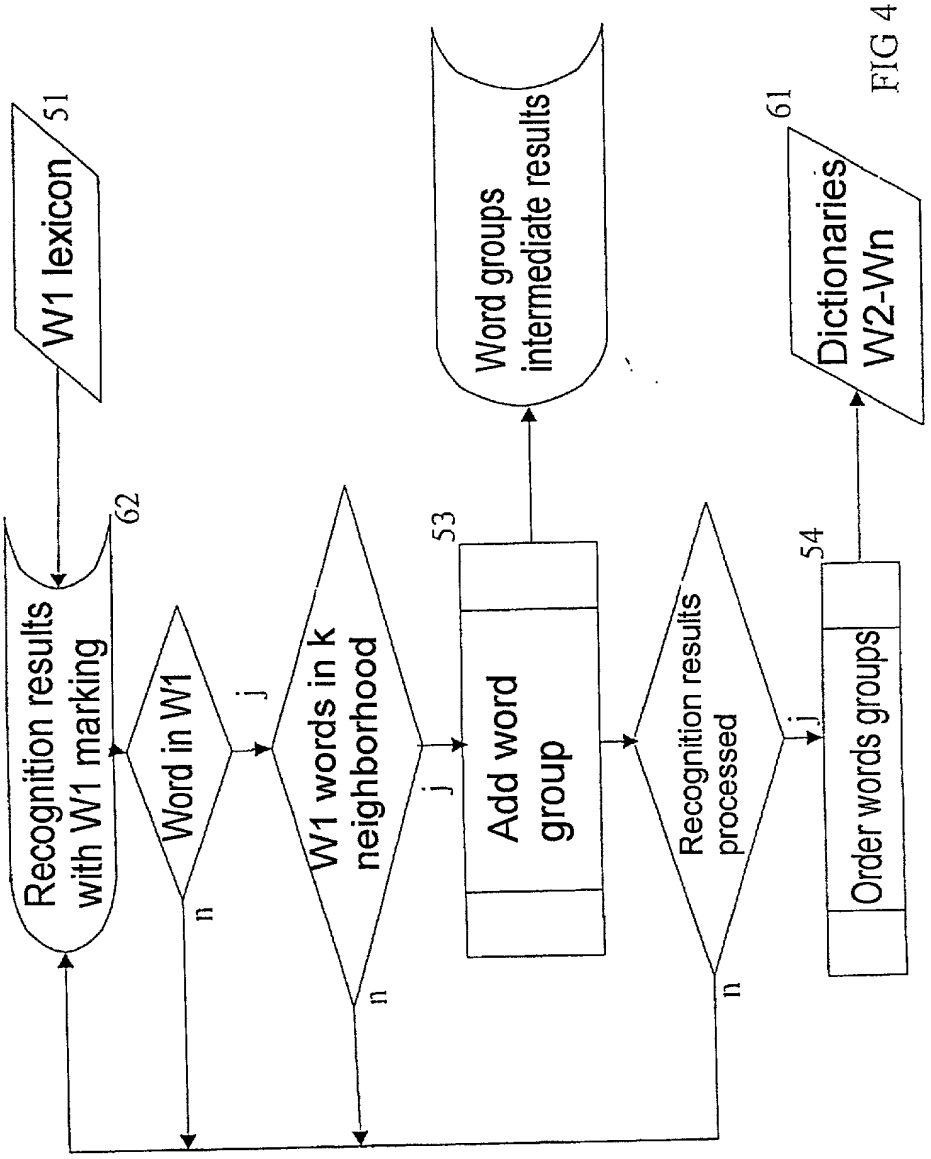

W1 lexicon
51
Recognition results with W1 marking
62
Word in W1
j
n
W1 words in k neighborhood
j
n
53
Add word group
Word groups intermediate results
Recognition results processed
j
n
61
Dictionaries W2-Wn
54
Order words groups

FIG 4

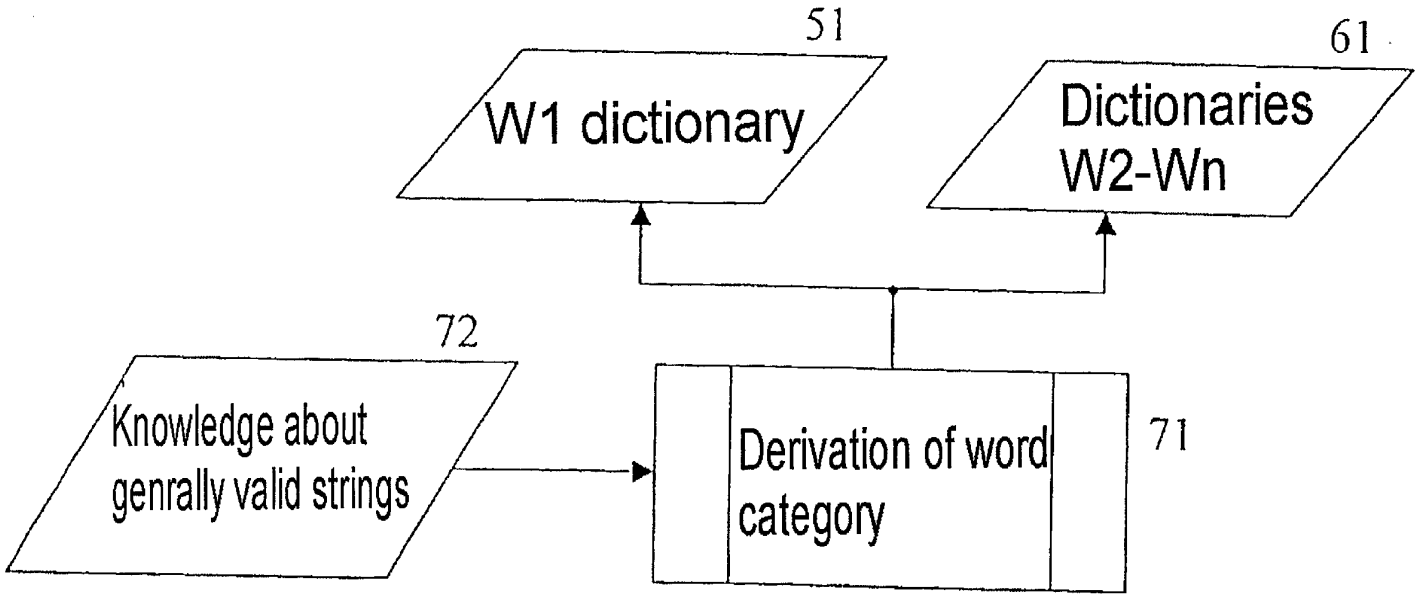


FIG 5

Declaration and Power of Attorney For Patent Application
Erklärung Für Patentanmeldungen Mit Vollmacht
German Language Declaration

Als nachstehend benannter Erfinder erkläre ich hiermit an Eides Statt:

dass mein Wohnsitz, meine Postanschrift, und meine Staatsangehörigkeit den im Nachstehenden nach meinem Namen aufgeführten Angaben entsprechen,

dass ich, nach bestem Wissen der ursprüngliche, erste und alleinige Erfinder (falls nachstehend nur ein Name angegeben ist) oder ein ursprünglicher, erster und Miterfinder (falls nachstehend mehrere Namen aufgeführt sind) des Gegenstandes bin, für den dieser Antrag gestellt wird und für den ein Patent beantragt wird für die Erfindung mit dem Titel:

Verfahren zur Bildung und/oder Aktualisierung von Woerterbuechern zum automatischen Adresslesen

deren Beschreibung

(zutreffendes ankreuzen)
☐ hier beigefügt ist.
☒ am 31.05.2000 als
PCT internationale Anmeldung
PCT Anmeldungsnummer PCT/DE00/01791
eingereicht wurde und am _____
abgeändert wurde (falls tatsächlich abgeändert).

Ich bestätige hiermit, dass ich den Inhalt der obigen Patentanmeldung einschliesslich der Ansprüche durchgesehen und verstanden habe, die eventuell durch einen Zusatzantrag wie oben erwähnt abgeändert wurde.

Ich erkenne meine Pflicht zur Offenbarung irgendwelcher Informationen, die für die Prüfung der vorliegenden Anmeldung in Einklang mit Absatz 37, Bundesgesetzbuch, Paragraph 1.56(a) von Wichtigkeit sind, an.

Ich beanspruche hiermit ausländische Prioritätsvorteile gemäss Abschnitt 35 der Zivilprozessordnung der Vereinigten Staaten, Paragraph 119 aller unten angegebenen Auslandsanmeldungen für ein Patent oder eine Erfindersurkunde, und habe auch alle Auslandsanmeldungen für ein Patent oder eine Erfindersurkunde nachstehend gekennzeichnet, die ein Anmeldedatum haben, das vor dem Anmeldedatum der Anmeldung liegt, für die Priorität beansprucht wird.

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name,

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled

Verfahren zur Bildung und/oder Aktualisierung von Woerterbuechern zum automatischen Adresslesen

the specification of which

(check one)
☐ is attached hereto.
☒ was filed on 31.05.2000 as
PCT international application
PCT Application No. PCT/DE00/01791
and was amended on _____
(if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations, §1.56(a).

I hereby claim foreign priority benefits under Title 35, United States Code, §119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

German Language Declaration

Prior foreign appplications
Priorität beansprucht

			Priority Claimed	
19933984.8	DE	20.07.1999	☒	☐
(Number) (Nummer)	(Country) (Land)	(Day Month Year Filed) (Tag Monat Jahr eingereicht)	Yes Ja	No Nein
			☐	☐
(Number) (Nummer)	(Country) (Land)	(Day Month Year Filed) (Tag Monat Jahr eingereicht)	Yes Ja	No Nein
			☐	☐
(Number) (Nummer)	(Country) (Land)	(Day Month Year Filed) (Tag Monat Jahr eingereicht)	Yes Ja	No Nein

Ich beanspruche hiermit gemäss Absatz 35 der Zivilprozessordnung der Vereinigten Staaten, Paragraph 120, den Vorzug aller unten aufgeführten Anmeldungen und falls der Gegenstand aus jedem Anspruch dieser Anmeldung nicht in einer früheren amerikanischen Patentanmeldung laut dem ersten Paragraphen des Absatzes 35 der Zivilprozeßordnung der Vereinigten Staaten, Paragraph 122 offenbart ist, erkenne ich gemäss Absatz 37, Bundesgesetzbuch, Paragraph 1.56(a) meine Pflicht zur Offenbarung von Informationen an, die zwischen dem Anmeldedatum der früheren Anmeldung und dem nationalen oder PCT internationalen Anmeldedatum dieser Anmeldung bekannt geworden sind.

I hereby claim the benefit under Title 35. United States Code. §120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, §122, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, §1.56(a) which occured between the filing date of the prior application and the national or PCT international filing date of this application.

PCT/DE00/01791 (Application Serial No.) (Anmeldeseriennummer)	31.05.2000 (Filing Date D, M, Y) (Anmeldedatum T, M, J)	anhängig (Status) (patentiert, anhängig, aufgegeben)	pending (Status) (patented, pending, abandoned)
(Application Serial No) (Anmeldeseriennummer)	(Filing Date D,M,Y) (Anmeldedatum T, M; J)	(Status) (patentiert, anhängig, aufgeben)	(Status) (patented, pending, abandoned)

Ich erkläre hiermit, dass alle von mir in der vorliegenden Erklärung gemachten Angaben nach meinem besten Wissen und Gewissen der vollen Wahrheit entsprechen, und dass ich diese eidesstattliche Erklärung in Kenntnis dessen abgebe, dass wissentlich und vorsätzlich falsche Angaben gemäss Paragraph 1001, Absatz 18 der Zivilprozessordnung der Vereinigten Staaten von Amerika mit Geldstrafe belegt und/oder Gefängnis bestraft werden koennen, und dass derartig wissentlich und vorsätzlich falsche Angaben die Gültigkeit der vorliegenden Patentanmeldung oder eines darauf erteilten Patentes gefährden können.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

German Language Declaration

VERTRETUNGSVOLLMACHT: Als benannter Erfinder beauftrage ich hiermit den nachstehend benannten Patentanwalt (oder die nachstehend benannten Patentanwälte) und/oder Patent-Agenten mit der Verfolgung der vorliegenden Patentanmeldung sowie mit der Abwicklung aller damit verbundenen Geschäfte vor dem Patent- und Warenzeichenamt: *(Name und Registrationsnummer anführen)*

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith. *(list name and registration number)*

Young & Thompson

And I hereby appoint

Customer No. 00466

Telefongespräche bitte richten an: *(Name und Telefonnummer)*

Direct Telephone Calls to: *(name and telephone number)*

Young & Thompson
(001) 703 521 22 97

Postanschrift:

Send Correspondence to:

Young & Thompson
745 South 23rd Street, Suite 200 22202 Arlington, VA
Telephone: (001) 703 521 22 97 and Facsimile (001) 703 685 05 73
or
Customer No. 00466

Voller Name des einzigen oder ursprünglichen Erfinders: Dr. THOMAS BAYER	Full name of sole or first inventor: Dr. THOMAS BAYER
Unterschrift des Erfinders Datum 12/12/01	Inventor's signature Date 12/12/01
Wohnsitz RADOLFZELL, DEUTSCHLAND	Residence ~~RADOLFZELL~~, GERMANY DEY
Staatsangehörigkeit DE	Citizenship DE
Postanschrift HOERIBLICK 10	Post Office Addess HOERIBLICK 10
D-78315 RADOLFZELL DEUTSCHLAND	D-78315 RADOLFZELL GERMANY
Voller Name des zweiten Miterfinders (falls zutreffend).	Full name of second joint inventor, if any:
Unterschrift des Erfinders Datum	Second Inventor's signature Date
Wohnsitz ,	Residence ,
Staatsangehörigkeit	Citizenship
Postanschrift	Post Office Address

(Bitte entsprechende Informationen und Unterschriften im Falle von dritten und weiteren Miterfindern angeben).

(Supply similar information and signature for third and subsequent joint inventors).